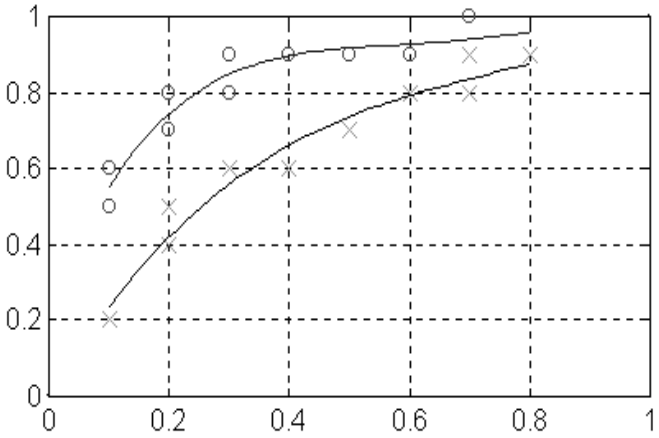
Testy detekcji oparte na symulacji i analizie statystycznej

Cechy:

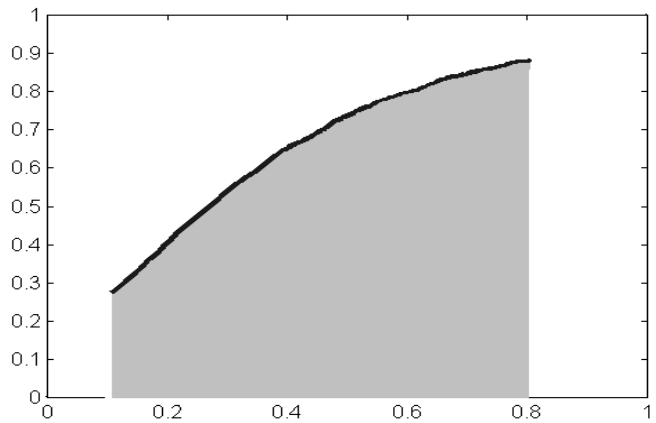
- częste wykorzystywanie krzywej ROC, stosowanie wielostopniowej skali ocen np. pięciostopniowej (pewna cecha jest zdecydowanie obecna, prawdopodobnie obecna, może obecna, prawdopodobnie nieobecna lub też definitywnie nieobecna),
- specjaliści są wcześniej przygotowywani do podejmowania decyzji w przełożeniu na skalę ocen,
- zapewnia się takie warunki testu, które z jednej strony jak najbardziej odpowiadają konkretnym warunkom pracy, z drugiej zaś ograniczają do minimum czynniki zakłócające obiektywną oceną (proces uczenia, skojarzenia, zmienne warunki obserwacji itd.),
- wyniki decyzji poszczególnych specjalistów podlegają następnie analizie statystycznej w celu wyznaczenia sumarycznych wskaźników wyrażających ocenę jakości obrazów.

Przykłady metod analizy krzywych ROC:

- aproksymacja punktów pomiarowych wielomianami, funkcjami sklejanymi z minimalizacją błędu średniokwadratowego itp.



- obliczanie pola pod krzywą



- testy statystyczne (weryfikacja hipotez statystycznych)

porównywać można wartości pól pod krzywymi dla każdego lekarza, parametry funkcji aproksymujących lub też dokładne wartości punktów testowych

<u>prosty test ze statystyką *U*</u>

Mamy: dwie duże, niezależne próby pobrane z populacji niekoniecznie normalnych, o nieznanych wartościach średnich m_1 i m_2 i o nieznanych, lecz równych wariancjach σ_1^2 i σ_2^2. Hipoteza: $H_0: m_1 = m_2$. Statystyka:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

której rozkład przy prawdziwości hipotezy H_0 jest asymptotycznie normalny *N(0,1),* n_1, n_2 - liczebności prób.

Reguła postępowania jest następująca:

- przybliżamy wartości średnich i wariancji poprzez $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$,
- ustalamy poziom istotności α
- rozpatrujemy, którą z hipotez alternatywnych należy wziąć pod uwagę:

$$H_1: m_1 \neq m_2, \quad H_2: m_1 > m_2, \quad H_3: m_1 < m_2$$

- jeśli wybieramy H_1, to stosujemy test dwustronny i **<u>odrzucamy hipotezę</u>** H_0 na korzyść hipotezy H_1, gdy dla obliczonej wartości

  - $$u_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

spełniona jest nierówność $|u_0| > u_\alpha$,

gdzie u_α jest wartością statystyki U wyznaczoną z tablicy rozkładu normalnego, dla której $P(|U| \geq u_\alpha) = \alpha$.

Jeśli hipotezą alternatywną względem hipotezy H_0 jest hipoteza H_2 to stosujemy test prawostronny i odrzucamy H_0 na korzyść H_2 jeśli $u_0 > u_\alpha$ itd.

t-Studenta

W tych zastosowaniach można użyć następującego testu:

weryfikujemy hipotezę, że średnie dwóch małych prób nie różnią się istotnie, czyli $H_0: m_1 = m_2$. Przyjmujemy, że próby są niezależne, a populacje w przybliżeniu normalne o nieznanych, lecz równych wariancjach (niestety niezbyt spełnione w tym przypadku). Podstawą testu jest statystyka:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Ma ona rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody

Można też stosować inne testy, np. F-Snedecora dotyczący weryfikacji hipotezy o równości wariancji empirycznych rozkładów, itp.

Ocena wiarygodności diagnostycznej obrazów medycznych

Cechami charakterystycznymi tych metod są przede wszystkim:

- duża złożoność i czasochłonność,
- wykorzystanie subiektywnych opinii lekarzy-specjalistów w danej dziedzinie, przy jednoczesnym dążeniu do maksymalnej obiektywizacji tych ocen,
- stworzenie warunków oceny jakości obrazów rekonstruowanych zbliżonych do codziennej praktyki lekarskiej.

Wady techniki ROC:

- konieczność zamiany normalnego trybu diagnozowania w praktyce klinicznej na wyrażenie opinii w pewnej skali ocen.
- ponieważ technika ROC została stworzona przy założeniu Gaussowskiego rozkładu szumów w zbiorze analizowanych danych, jej stosowanie do oceny danych obrazowych o zazwyczaj nie-Gaussowskim charakterze nasuwa pewne wątpliwości

(istnieją pewne metody redukcji błędów wynikających z tych Gaussowskich założeń).

- wiele praktycznych zadań diagnostycznych, jakie stoją przed specjalistami, nie sprowadza się do decyzji dwupoziomowej: tak lub nie; w niektórych patologiach występuje kilka nieprawidłowości w różnych fragmentach obrazu, a proces decyzyjny jest dużo bardziej złożony.

Modyfikacje ROC

- Specjaliści obserwują obrazy i zaznaczają obecność pewnych anormalności np. powiększonych węzłów chłonnych w obrazie CT klatki piersiowej, lub też obecność guzków w płucach, przy czym liczba anormalności jest różna w poszczególnych obrazach testowych.
- Warstwa decyzyjna zostaje rozszerzona na kilka, czy nawet kilkanaście poziomów.
- Analiza tak otrzymanych wyników przeprowadzana jest przy pomocy dwu parametrów: czułości i przewidywanej wartości pozytywnej PVP (*predictive value positive*) - $\frac{N_{pp}}{N_{pp} + N_{fp}}$.

Jeśli obserwator zakreśli wszystkie anormalności w obrazie, wówczas osiąga maksymalną wartość czułości 1, a jeśli mniej - odpowiedni ułamek wyraża czułość jego decyzji. Natomiast PVP określa szansę rzeczywistej obecności anormalności w zaznaczonych miejscach. Jeżeli więc ekspert byłby zbyt agresywny w wykrywaniu anormalności, wówczas dużej wartości czułości będzie towarzyszyć mała wartość PVP, a w przypadku zbytniej ostrożności wyniki będą dokładnie odwrotne. Następnie, na wykresach przedstawiane są średnie wartości czułości i PVP (oddzielnie) dla każdego stopnia kompresji badanych obrazów, które aproksymuje się odpowiednią funkcją, np. kwadratową funkcję sklejaną z kryterium minimalizacji błędu średniokwadratowego. Pojedynczy punkt na wykresie odpowiada decyzji jednego specjalisty dla obrazu o danym stopniu kompresji.

- Porównanie czułości i PVP dla różnych stopni kompresji (wyznaczenie statystycznie istotnej różnicy ocen jakości) przeprowadzono przy pomocy testu t-Studenta z wykorzystaniem rozkładu permutacji dwuelementowych (nazywanego czasami testem Behrensa-Fishera)

Test ten nadaje się do danych, które nie mają gaussowskiego charakteru.

Załóżmy: specjalista 1 określa jakość *N* obrazów należących do dwóch poziomów A i B (np. obrazy kompresowane w dwu różnych stopniach). Obrazy te należą do dziewięciu grup: bez patologii, z jedną patologią, z dwoma patologiami, ..., z ośmioma patologiami. Przez N_i oznaczmy liczbę obrazów *i*-tej grupy, a Δ_{ij} niech reprezentuje różnicę wartości czułości (lub PVP) dla *j*-tego obrazu *i*-tej grupy oglądanego na dwóch poziomach jakości.

Niech $\overline{\Delta}_i$ będzie średnią różnicą według równania:

$$\overline{\Delta}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\Delta_{ij} \ .$$

Teraz **zdefiniujmy**:

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(\Delta_{ij} - \overline{\Delta}_i)^2,$$

a następnie ***t* statystyka Behrensa-Fishera** dana jest przez równanie:

$$t_{BF} = \frac{\sum_{i=1}^{N_i}\overline{\Delta}_i}{\sqrt{\sum_{i=1}^{N_i}\frac{S_i^2}{N_i}}}$$

Dla każdego z N obrazów można liczyć wartości testu z dwóch poziomów ($\mathrm{A} \to \mathrm{B}$ i $\mathrm{B} \to \mathrm{A}$) lub nie (pełna liczba permutacji wynosi 2^N). Obliczenia t_{BF} wykonywane są dla całego rozkładu permutacji. Jeśli k jest liczbą wartości t_{BF}, które przekraczają wartość 'prawdziwą' (przypadek: ten sam obraz z poziomu A i B), wtedy wartość $\frac{(k+1)}{2^N}$ jest przyjmowana jako poziom istotności testów jednostronnych zerowej hipotezy, że jakość obrazów dla wyższego stopnia kompresji jest przynajmniej tak dobra jak obrazów niższego stopnia kompresji.

Nowa metoda bez ROC:

- określenie złotego standardu: zgodny, osobisty, niezależny i osobny
- obiektywna diagnoza powstaje na podstawie analizę obrazów analogowych, względem których ocenia się obrazy cyfrowe (także cyfrowy oryginał)
- nowy protokół oceny jakości obrazów, w którym lekarze wyrażają swe opinie w kategoriach jak najbardziej diagnostycznych
- analiza statystyczna bez krzywych ROC:

do zapisu wyników testu wykorzystano tablicę 2×2 postaci

II\I	**dobrze**	**źle**
dobrze	**N(1,1)**	**N(1,2)**
źle	**N(2,1)**	**N(2,2)**

gdzie I - może oznaczać obraz oryginalny analogowy a II - oryginał cyfrowy lub też I - oryginał cyfrowy, II - cyfrowy kompresowany. Decyzje **dobrze** znaczą zgodnie ze złotym standardem, a **źle** - niezgodnie.

Jeśli uzyskane tablice nie są diagonalne, to do oceny statystyki uzyskanych wyników stosuje się test McNemara:

porównanie N(1,2) i (N(1,2)+N(2,1))/2 z pomocą rozkładu dwumianowego o parametrach N(1,2)+N(2,1) i 0.5

lub Fishera:

liczymy statystykę $\left(N(1,2) - N(2,1)\right)^2 / \left(N(1,2) + N(2,1)\right)$, która ma rozkład chi-kwadrat z jednym stopniem swobody.

Zastosowano także tablice zgodności decyzji radiologów:

	RTS	F/U	C/B	BX
RTS	12	0	5	0
F/U	0	0	0	0
C/B	3	0	12	6
BX	0	0	2	17

RTS - przypadkowy, negatywny, lub łagodny do powtórnego badania,

F/U - prawdopodobnie łagodny ale wymagający sześciomiesięcznej obserwacji,

C/B - potrzebne dodatkowe badania,

BX - biopsja.

Porównano badania analogowe i dyskretne.

rozwiązania te są stosowane dla oceny jakości obrazów mammograficznych.

Perspektywa dalszych badań:

- możliwie pełne wprowadzenie zrekonstruowanych obrazów w rzeczywiste warunki pracy klinicznej i wnioskowania o ich jakości na podstawie testów, które coraz silniej naśladują realną ocenę diagnostyczną prezentowanych struktur.
- określenie **złotego standardu** jako obiektywnej diagnozy.

Złoty standard wyraża 'prawdę' diagnostyczną każdego oryginalnego obrazu (wykorzystywanego w teście) i służy jako baza do porównań wyników diagnoz dokonywanych na podstawie różnych wersji (po kompresji) tego obrazu. Złoty standard może być wyznaczany na wiele sposobów (standard zgodny, osobisty, niezależny, osobny), w zależności od tego, co tak naprawdę powinien wyrażać; czy osobiste przekonanie pojedynczego lekarza, czy też pewien kompromis pomiędzy kilku specjalistami oceniającymi później także obrazy rekonstruowane lub też zupełnie niezależnymi od późniejszych ocen. Koncepcja standardu osobnego proponuje w celu sformułowania prawdy obiektywnej o obrazie oryginalnym skorzystanie z innych badań (chirurgicznej biopsji, innych badań obrazowych), obserwacji pacjenta itd.

- ocena jakości oryginału, a nie tylko obrazu rekonstruowanego (np. względem oryginału analogowego),
- usprawnienie protokołu w testach psychowizualnej oceny jakości obrazów,
- rezygnacja z ROC na korzyść bardziej przydatnych metod statystycznej analizy wyników,
- konstrukcja coraz doskonalszych modeli systemu HVS,
- tworzenie obliczeniowych miar jakości o cechach całkowo-różniczkowych (są to najczęściej miary wektorowe).